

How to dissect surgical journals: IX – Sample size*

The difference between observed values from a sample and the true value diminishes as the number of observations increases.

Law of Large Numbers, Jacob Bernoulli (1654–1705)

The third biggest lie in the universe is that ‘P values’ measure the strength of relationships. It pervades the clinical literature and subverts the minds of unwary readers. Let me illustrate. The following table details three comparisons, each comparing proportions of 10% and 20%. In other words, the magnitude of the difference between the groups – what biostatisticians call the ‘effect size’ - is fixed. The comparisons were made using two-tailed chi-square tests with Yates’s continuity correction:

1/10	v	2/10	P = 0.53 (not significant)
10/100	v	20/100	P = 0.05 (borderline significance)
100/1,000	v	200/1,000	P < 0.001 (significant)

The comparison becomes increasingly significant as the sample size increases, which confirms that:

- P values are a measure of the reliability of the observed differences.
- The reliability of the results increases as the sample size increases.

Look what happens when we freeze the sample size and increase the differences between the groups:

1/10 (10%)	v	3/10 (30%)	P = 0.26 (not significant)
1/10 (10%)	v	5/10 (50%)	P = 0.05 (borderline significance)
1/10 (10%)	v	7/10 (70%)	P = 0.006 (significant)

This demonstrates that it is easier to reliably detect large differences than small differences. Failure to appreciate these simple relationships leads to confusion. As mentioned in the previous chapter, many articles skip over the observed differences between the groups and concentrate on the results of statistical tests. The problems are that

*The sixteen articles in this series are being made available on ANZJSurg.com as an eBook.

clinically useful differences in small populations may fail to achieve statistical significance; and, small differences in large populations may be statistically significant but be clinically irrelevant.

Big numbers smooth out variations due to chance, but they cannot overcome systematic errors (bias). To paraphrase some of my previous comments:

Selection procedures determine accuracy.
Sample size determines precision.

The fundamental attribution error

The problem of ‘variations due to small numbers’ has a social as well as a statistical dimension. Lee Ross¹ described this as ‘the fundamental attribution error’. We tend to think that outcomes are due to individuals rather than the circumstances. Individuals are made the scapegoats for systemic failures and are praised for the success of organisations, even when the circumstances made success inevitable. James Surowiecki² used the ‘fundamental attribution error’ to explain how the introduction of the Boeing 787 led commentators to attribute Boeing’s success to skilful management, rather than fortuitous timing; and, at the same time, to predict the imminent demise of Airbus despite its sustained period of market leadership.

This type of problem occurs in the surgical literature. We underestimate how much variation can be caused simply by luck and see patterns where none exist. This leads us to be overtly influenced by the nature of the message rather than the soundness of the study. It is a part of the ‘confirmation bias’ that was described in the first chapter.

Nothing defines humans better than their willingness to do irrational things in the pursuit of phenomenally unlikely payoffs.

Scott Adams (1997)

Empiricism dictates that progress is made through experience. Ideas that initially sound reasonable can fail to progress; and sometimes, in retrospect, they are judged to be foolish. In the broad scope of things, this is an essential process because it leads to advances. But it presents a hazard for those at the cutting edge.

Table 1 The errors that can occur when comparing two groups

	Probability of detecting a difference	Probability of not detecting a difference
No difference between the populations	Alpha (α) (Type I Error) (False-Positive)	1 – Alpha (α) (Confidence) (True-Negative)
A difference between the populations	1 – Beta (β) (Power) (True-Positive)	Beta (β) (Type II Error) (False-Negative)

False-negative results (type II errors)

There are two aspects to the unreliability of small numbers – ‘false positives’ and ‘false negatives’. The last article in this series provided a framework for understanding the risk of false-positive results (type I errors). The other side of the coin is the possibility of false-negative conclusions i.e. studies that claim that there is no appreciable difference in an outcome when there may be (type II errors). As Table 1 indicates, the probability of a type II errors is referred to as beta (β).

The confidence that you can have in the results of a study is influenced by the sample size. It is impossible to have faith in ‘negative results’ derived from studies that have used inadequate sample sizes.

It is a common and serious error to conclude that ‘no effect exists’ just because the P value fails to achieve statistical significance

There is a concern that many interventions that are labelled as being ineffective have not received a fair test.

Estimating the required sample size

Estimates of the sample size are based on:³

- Alpha (the probability of a type I error),
- Beta (the probability of a type II error),
- The response rate for the control group, and
- Delta (the minimum important difference).

The probability of a type I error (alpha) of 5% is easy to fix into the equation – there is near universal agreement that $P > 0.05$ constitutes ‘statistical significance’ for a two-tailed comparison. You might therefore believe that the probability of a type II error (beta) of 5% would be appropriate, and you would be right if the clinical literature were equally averse to making false-positive and false-negative conclusions. However, convention sets it at a much higher figure, and it is easy to understand why.

The number of clinical trials published in the medical literature would fall dramatically if they were required to recruit enough patients to satisfy the requirements of a power ($1-\beta$) of 95%. Instead,

it has been silently agreed that false-negative errors are not as important as false-positive errors. Power is usually set at 80% or 90%. Also, authors, if they bother to raise the issue at all, usually mention ‘power’ rather than the probability of type II errors. This reflects a bias towards detecting a difference (power & alpha) rather than missing a difference (confidence & beta).

The last group of numbers relate to the anticipated response rates. Hopefully, the investigators will have a reasonable estimate of the response rate in the control group from pilot studies and a study of the literature. They then need to declare a value for delta (the minimum important difference).

The minimum important difference

The ‘minimum important difference’ is the smallest difference in outcome that would be of either biological or clinical interest. For us, this is the difference that would alter our surgical practice. There is no formula or edict from a higher source that dictates the importance of any given difference. It is a value judgement.

Determining a value for the minimum important difference is often subjective. It is influenced by the seriousness of the condition, the outcome being measured, the importance of side-effects, costs and inconveniences, and the availability of alternative treatments.⁴ What is important in one clinical setting may not be important in another. For example, at the end of a trial the incidence of the outcome event might be 2% versus 4%. A trial evaluating prophylactic antibiotics in clean surgery might observe that the incidence of wound infection is halved in the intervention group. Which is no doubt a good thing; but this has to be weighed against the side effects and the total costs of administering the antibiotic. On the other hand, cutting the risk of death in half after a major elective procedure would be of considerable interest.

Kashani *et al.*⁵ found that only 21% (100/486) of admissible surgical trials mentioned a value for the minimum clinically important difference when estimating the sample size. They estimated that ‘one-third of the surveyed trials needed to accrue at least fivefold the number of patients, which, in pragmatic terms, means that many trials are doomed from the outset, have no prospect of providing a reliable result, and have squandered valuable clinical and financial resources’. Nevertheless, there are plenty of examples within the surgical literature.

Heal *et al.*⁶ evaluated whether a single application of topical chloramphenicol would reduce the incidence of wound infection after minor surgery. They based their sample size on anticipated infection rates of 10% (control) versus 5% (chloramphenicol) based on one of their previous clinical trials i.e., a minimum important difference of 5%. Their observed response rates were 11.0% versus 6.6% ($P = 0.01$), a difference of 4.4%. They concluded that topical chloramphenicol produced ‘a moderate absolute reduction in infection rate that is statistically but not clinically significant’. Another interesting point was their use of paraffin ointment as a placebo control. This was important because the clinical staff performing the surgery also assessed the wounds. They could not get ‘information about the exact proportions of the constituents of the base of Chloramphenicol ointment from the manufacturer. The principal investigator visited a compounding pharmacist to develop a close match to the vehicle of

Table 2 Examples from Fleiss *et al.*⁷ of the sample size required *in each group* when using a two-tailed test with the probability of a Type I Error set at 5% and a response rate of 10% in one of the groups. Delta is the difference in response rates between the groups

Response rates	Delta	70% power	80% power	90% power	95% power
15% v 10%	5%	579	725	957	1,174
20% v 10%	10%	176	219	286	348
25% v 10%	15%	91	113	146	177
30% v 10%	20%	58	71	92	111

the Chloramphenicol ointment by using a mixture of soft white and liquid paraffin, prepared single doses of the ointment in sterile jars, and stored them in a refrigerator . . . Only the principal investigator was aware of the identity of the coded ointments’.

Number crunching

Outcomes can be expressed as proportions, continuous measurements, or survival curves. It is possible to estimate the required sample size for each of these circumstances. When estimating the sample size for continuous data, a large standard deviation indicates considerable variability and, hence, the need for a relatively large sample size. Tables are available for rates and proportions (Table 2). When estimating the sample size for survival curves, small gaps between the lines indicates the need for a relatively large sample size.

The convention is to use ‘N’ for populations and ‘n’ for samples

Investigators have to live within the bounds of their available resources. What happens if they estimate the required sample size and it is impossible for them to collect that many patients? Do they abandon the study? Maybe, but in practice they often produce figures based on ‘the patients that I can get’, or more often perhaps on ‘the patients that I got’, rather than on ‘the patients that I need’. So, in practice, the declared sample size is often a compromise between what is useful and what is achievable.

Investigators often bend the rules when reporting sample sizes. Chan *et al.*⁸ compared the protocols for clinical trials approved by an institutional ethics committee with the subsequent publications. They concluded that: ‘When reported in publications, sample size calculations and statistical methods were often explicitly discrepant with the protocol or not pre-specified. Such amendments were rarely acknowledged in the trial publication. The reliability of trial reports cannot be assessed without having access to the full protocol’. So, it is useful for interested readers to have access to registers of trials.

How can investigators drop the numbers of patients that are required? Let’s take the example of anticipated response rates of 20% v 10% (i.e. a minimum important difference of 10%), with alpha set at 0.05 using a two-tailed comparison, and a power of 80%. The required sample size is 219 patients per group. Choosing to use

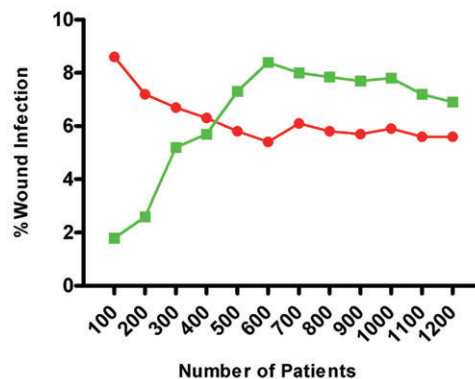


Fig. 1. The cumulative rate of wound infection in a study evaluating the use of antimicrobial agents in patients undergoing abdominal surgery.⁹

a one-tailed comparison drops it down to 176 patients. If the power is then dropped to 70% it becomes 139 patients. And, if, in addition, the minimally important difference is re-estimated at 20% the estimated required sample size becomes 47 patients per group. So, the total sample size drops from 438 to 94 patients. As an interested reader you need to do more than just check that the investigators have declared a formal calculation of the sample size. You need to see the details. Your mantra must be: ‘show me the numbers’.

Finally, another way to decrease the estimated sample size is to collect more information – statistically, it is the amount of information that is important not just the number of patients. The power of a study can be increased by making more measurements e.g., the use of multiple endpoints rather than just one declared event, recording measurements over time rather than just at the end of the study.

The importance of monitoring

In an earlier chapter, I discounted the ‘intraocular trauma test’ as a way of evaluating results. But sometimes trends can be informative. Some time ago I was stumbling around the use of sequential analysis to analyze clinical trials. What caught my attention was the usefulness of graphing the cumulative outcome over the course of a study. I did this at the end of a clinical trial that compared two antibiotics in patients undergoing abdominal surgery. Figure 1 demonstrates a couple of things. First, there would have been a significant difference between the agents if the trial had stopped after evaluating a couple of hundred patients. Second, the trial results achieved stability – as indicated by the lines starting to run parallel – after evaluating about 800 patients. It is an iconic image.

Some further considerations

Be wary of:

- Comparative studies that use P values to claim equivalence between the groups at the start of a study. Testing for baseline equivalence, if it was ever appropriate, would require a consideration of the probability of false-negative (type II) errors.
- Claims based on small sample sizes, regardless of the attractiveness of the message. The surgical literature is biased towards the reporting of positive results.
- Reported estimates of the required sample size – especially if the aim of the trial is to demonstrate that there is no appreciable difference between the interventions i.e. an ‘equivalence trial’. Here, it is reasonable to expect the investigators to employ a power of 95%.

No amount of sophisticated ‘black-box’ number-crunching can compensate for a small sample size. Type ‘sample size calculation’ into your favourite search engine if you wish to play around with some numbers.

Key Points

- The reliability of results increases as the sample size increases.
- It is easier to reliably detect large differences than small differences.
- Type II errors occur when it is falsely claimed that there is no appreciable difference in an outcome when there is (false-negative).
- It is wrong to conclude that ‘no effect exists’ just because the P value fails to achieve statistical significance.
- Estimates of the sample size are based on: α , β , the response rate for the control group, and Delta (the minimum important difference).
- The minimum important clinical difference is the smallest difference in outcome that would be of clinical interest i.e., it might alter your clinical practice.
- When estimating the sample size for continuous data, a large standard deviation indicates considerable variability and, hence, the need for a relatively large sample size.

- Tables and nomograms are available to calculate sample sizes for rates and proportions.
- Investigators often bend the rules when reporting sample sizes.
- Be wary of claims based on small sample sizes, regardless of the attractiveness of the message.

References

1. Ross L. *The intuitive psychologist and his shortcomings: Distortions in the attribution process*. In Berkowitz L. (Ed.), *Advances in Experimental Social Psychology*. Academic Press: New York, 1977.
2. Surowiecki J. The fatal-flaw myth. *The New Yorker*, 31 July, 2006.
3. Machin D, Campbell M, Fayers P, Pinol A. *Sample Size Tables for Clinical Studies*. 2nd edition. Oxford: Blackwell Sciences, 1997.
4. Man-Son-Hing M, Laupacis A, O’Rourke K *et al*. Determination of the clinical importance of study results. *Rev. J. Gen. Intern. Med.* 2002; **17**: 469–76.
5. Kashani I, Hall JL, Hall JC. The estimation of minimum clinically important differences in surgical trials. *ANZ J. Surg.* 2009; **79**: 301–4.
6. Heal CF, Buettner PG, Cruickshank R *et al*. Does single application of topical chloramphenicol to high risk sutured wounds reduce incidence of wound infection after minor surgery? Prospective randomised placebo controlled double blind trial. *B.M.J.* 2009; **338**: a2812.
7. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions*, 3rd edition, Wiley-Interscience: New York, 2003.
8. Chan A-W, Hróbjartsson A, Jørgensen KJ, Gøtzsche PC, Altman DG. Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *B.M.J.* 2008; **337**: 1404–7.
9. Hall JC, Hall JL, Christiansen K. The role of ceftriaxone and cephalexin in patients undergoing abdominal surgery: A clinical trial. *Arch. Surg.* 1991; **126**: 512–6.

John C. Hall, MS, DS, FRACS
 Surgery (Royal Perth Hospital),
 The University of Western Australia,
 Perth, Western Australia, Australia
 john.hall@uwa.edu.au

doi: 10.1111/j.1445-2197.2010.05359.x